
# CRESC Working Paper Series

Working Paper No. 138

## Socialising Big Data: From concept to practice

Evelyn Ruppert, Penny Harvey, Celia Lury, Adrian Mackenzie, Ruth McNally, Stephanie Alice Baker, Yannis Kallianos, Camilla Lewis

CRESC, The University of Manchester and the Open University

February 2015


# Socialising Big Data: From concept to practice

**Evelyn Ruppert[a], Penny Harvey[b], Celia Lury[c], Adrian Mackenzie[d], Ruth McNally[e], Stephanie Alice Baker[a], Yannis Kallianos[b], Camilla Lewis[b]**

[a] Goldsmiths, University of London
[b] CRESC, University of Manchester
[c] Centre for Interdisciplinary Methodologies, University of Warwick
[d] University of Lancaster
[e] Anglia Ruskin University

## Abstract

The working paper is a report on an ESRC-funded project, Socialising Big Data, that sought to address problematic conceptions of Big Data in popular discourse such as the 'data deluge' and the tendency to reduce the term to definitions such as the oft-cited '3 Vs'. Instead, building on how social scientists have conceived of things, methods and data as having social and cultural lives, the project sought to identify the normative, political and technical imperatives and choices that come to shape Big Data at various moments in its social lives. Recognising that Big Data involves distributed practices across a range of fields, the project experimented with collaboratories as a method for bringing together and engaging with practitioners across three different domains – genomics, national statistics and waste management. In this way it explored how relations between data are also simultaneously relations between people and that it is through such relations that a shared literacy and social framework for Big Data can be forged.

# Table of Contents

## Section 1: Introduction

### How we started thinking about this topic

*Socialising Big Data: Identifying the risks and vulnerabilities of data-objects* was an Economic and Social Research Council (ESRC) funded project that took place from June 2013 to Sept 2014. Our interdisciplinary collaboration involved a team of social scientists from a range of backgrounds (sociology, anthropology, and science and technology studies), many of whom were affiliated with the Centre for Research on Socio-Cultural Change (CRESC Manchester and The Open University) and the Centre for Economic and Social Aspects of Genomics (CESAGEN Lancaster), but also including other institutions.[1]

Our project aimed to advance the social scientific analysis of Big Data and digital practices to benefit academics, students, practitioners and policy makers. It emerged in response to the contemporary turn to Big Data in business, government and academia, and the idea that this topic was not well defined or understood. Our proposal highlighted problematic conceptions of Big Data in popular discourse, including the 'data deluge' and the tendency to reduce the term to definitions based on the '3 Vs': the increasing *volume* of data sets, *velocity* of data generation, and *variety* of data sources and formats.[2] While we recognised that these qualities make data more difficult to analyse using traditional management and processing applications, we highlighted that Big Data is not simply *generated by*, but also *generative of* innovations in computational and processing tools and analytics as well as novel ways of measuring and knowing phenomena.

Consequently, rather than attempting to define Big Data according to generic qualities (e.g. volume, velocity and variety), we aimed to focus on the specific *socio-technical practices* through which data is generated (e.g. online activities, mobile phone use, commercial and government transactions, sensors, sequencers and crowdsourcing), interpreted and made meaningful for analysis (e.g. mined, cleaned, linked, analysed, interpreted, stored and curated). From this perspective the challenge of Big Data is not simply its volume, but that working with Big Data creates new problems, risks and vulnerabilities given the tendency to overlook the social

[1] PI: Evelyn Ruppert, Goldsmiths, University of London. Co-Is: Penny Harvey, Manchester, CRESC; Celia Lury, Centre for Interdisciplinary Methodologies, Warwick; Adrian Mackenzie, Lancaster; Ruth McNally, Anglia Ruskin. Researchers: Stephanie Alice Baker, Goldsmiths, University of London; Yannis Kallianos and Camilla Lewis, University of Manchester, CRESC.

[2] Stapleton, L. K. (2011). Taming Big Data. *IBM Data Magazine*. 16: 1-6.

lives of data, which are neither natural nor technical phenomena, but enacted through multiple, selective social and technical practices. Our project, thus, sought to understand the often-unacknowledged normative and political effects of Big Data by investigating how methodological, digital and analytical practices enact and govern social worlds, of not only what is represented but also realised.

This approach and understanding are captured in the title of this project, Socialising Big Data. Picking up from how social scientists have conceived of things, methods and data as having social and cultural lives[3], we started by thinking about Big Data as having 'lives' that include social and technical practices that bring them into being (generate) but also order, manage, interpret, circulate, reuse, analyse, link and delete them. For each of these practices we sought to inquire into the normative, political and technical imperatives and choices and the actors and institutions that come to shape Big Data at various moments in their social lives.

This understanding necessitated the development of a method that would enable us to investigate the specificities of practices as they are being done and understood in particular contexts. But rather than doing this through discursive, ethnographic, interview or observational methods, we sought to experiment with a form of participatory research. We contended that by working *collaboratively* with practitioners in three domains – genomics, national statistics and waste management - rather than in our conventional separate roles as researchers, informants and users of research, we could co-produce shared concepts and understandings about Big Data that would be of use to diverse academic and non-academic stakeholders. We approached **collaboratories** as a model for doing this: *a collective, socialised method for identifying shared problems, concepts, and findings through synthetic, recursive engagements*. It is a model that has affinities with other experiments such as a research theme of UCI's Center for Ethnography where events such as seminars and workshops are understood as 'para-sites,' that is, as integral and designed parts of fieldwork that combine research, reflection and reporting and

---

[3] Kopytoff, Igor (1986). 'The Cultural Biography of Things: Commoditization as Process,' in *The Social Life of Things*, ed. Arjun Appadurai. Cambridge University Press, 64-91.

Lash, Scott and Celia Lury (2007). *Global Culture Industry: The Mediation of Things*, Cambridge, Polity.

Law, John, Evelyn Ruppert and Mike Savage (2011). 'The Double Social Life of Methods,' CRESC Working Paper Series, Paper No. 95.

Beer, David, and Roger Burrows (2013). 'Popular Culture, Digital Archives and the New Social Life of Data,' *Theory, Culture & Society* 30, 4: 47-71.

involve a mix of participants from the academy and specific communities of interest.[4] Understood as an overlapping academic/fieldwork space, para-sites exist outside conventional notions of the field and involve testing and developing ideas with communities not as key informants but as collaborators. For the Socialising Big Data project, we did this by organising our project around a series of events that experimented with and tested ways of engaging social scientists and practitioners in collaborative discussions and the co-production of concepts for understanding Big Data. Given our aim to 'socialise' Big Data, concept development formed an integral part of our approach.

### How we organised the project conceptually

In what follows, we provide a brief summary of key concepts that initially informed our interdisciplinary and collective team approach to, and development of, the collaboratories. The aim of these summaries was not to arrive at settled definitions, but rather to outline key concepts and indicative readings in the social sciences, which could serve as a conceptual starting point. Additionally, our purpose was not to subject these to definitional debates in the collaboratories, but instead to translate them into a series of specific questions and provocations that would enable us to revisit and revise them. With this in mind, we identified five concepts at the outset: Digital Data, Big Data, Digital Data Object (DDO), Boundary Object, and Collaboratories. These are briefly noted here and summarised in more detail in the Appendix.

Our initial object of analysis was what is commonly referred to as Big Data. Initially, we related this to understandings of the empirical turn to new forms of **Digital Data** more generally. Here we sought to reflect upon the ubiquity of digital devices and the data they generate – from social media platforms and browsers to online purchasing and sensors – and their implications for empirical methods in the social sciences. We noted that while these platforms and data are (usually) configured and owned by commercial actors and thus represent a challenge to the knowledge-making authority of social scientists, they also present an opportunity to rethink social science and other methods in ways that are 'closer' to social worlds and provide a provocation to invent methods that can adapt, re-purpose and engage with digital media in new and lively ways. In this regard, we sought to situate practitioner dilemmas and challenges in relation to social scientific ones.

---

[4] The Center for Ethnography (2009). 'Center as Para-site in Ethnographic Research Projects', University of California, Irvine. http://www.socsci.uci.edu/~ethnog/theme3.htm. As the Center's website notes, the term 'para-sites' was inspired Marcus, George E. (ed.) 2000. *Para-Sites: A Casebook Against Cynical Reason*, Chicago, University of Chicago Press.

Within this context our objective was not to define **Big Data**, but to focus on the specific situated practices through which such data is being generated and analysed (e.g. how data is captured, formatted, curated, stored, searched, traced, linked, shared, visualised). The diverse and far-reaching take up of the term across disciplines is indicative of the fundamental impact that Big Data is having from claims that it is reinventing society to inquiries about how it is changing the very material of scientific inquiry and knowledge and leading to alternative social theories of individuals and societies. While the 3 Vs has become the default definition in these domains, we turned our attention away from identifying qualities to investigating the social lives of Big Data by attending to practices to argue that what is 'big' about Big Data are novel ways of data generation, analysis, interpretation and implementation.

To do this, we initially experimented with specifying Big Data as a **DDO (digital data object)**. We drew the term from information and computing sciences where it is used to denote digitally stored data. However, we modified the term by using a designation from actor network theory – that of the data-object – to capture the network of practices and relations invested in its generation, maintenance and mobility. Through this conceptualisation, we sought to 'socialise' the DDO by attending to the interconnected and interdependent practices involved in generating and maintaining data (e.g., its detection, duration and deletion). While the term proved useful in capturing this relationality in a way that the term 'data' generally does not, it introduced two key problems: first, it implies an ontological differentiation between the subject and object (thereby instilling agency only with the former and not the latter), and second, the term has a very specific meaning in computing and information sciences, which is very much object-oriented.

The notion of the **boundary object** enabled us to consider the variety of ways that Big Data, whether understood as DDOs or not, are defined and conceived across communities of practice, between the highly technical and more general. While the meaning of boundary objects is malleable and varies in different social worlds, their structure is common enough to make them recognisable and useful across multiple sites (e.g. being weakly structured in common use, and strongly structured in individual-site use). Boundary objects are thus classifications that manage the tension between multiple interpretations across contexts where multiplicity is given and not incidental and are key to developing and maintaining coherence across intersecting social worlds.[5] From this perspective, Big Data is not a fixed object

[5] Bowker, G. C. and S. L. Star (1999). *Sorting Things Out: Classification and its Consequences.* Cambridge, Massachusetts, The MIT Press.

marked by certain qualities; rather, the same data is constituted and enacted in varying ways through different practices. One of the benefits of conceiving of Big Data as a boundary object is that the term captures the way in which objects emerge in relation to different communities of interest. Through specific situated practices particular definitions, problematisations and engagements with Big Data are constituted, each generating to different forms of uncertainty, risks and vulnerabilities.

We then approached **collaboratories** as a way to open up and engage practitioners with these concepts through a series of questions and provocations. In the social sciences, collaboratories include practices such as participatory research and partnerships with non-academic groups that seek to produce 'collective' rather than 'collected work'.[6] The benefit of this approach is that 'co-laboratories' mimic a laboratory in the sense that they favour an attitude of openness and experimentation. In the social sciences, the term 'collaboratory' has been adopted to capture interdisiciplinarity and working together with a wide range of collaborators. Inspired by the work of the Anthropology of the Contemporary Research Collaboratory (ARC), and recognising that it is not without its problems especially in terms of implementation, we took up this approach as a starting point for identifying some key features of the collaboratory as a method. In contrast to the ARC, however, we included practitioners, as well as academics, as co-producers of knowledge.

### How we organised the project practically

In practical terms, we were motivated to work with practitioners from a variety of contexts with different knowledge and expertise on and understandings of Big Data. Rather than reiterating the need to respond to the 'data deluge,' we sought to develop a *shared literacy* about Big Data by locating the successes and failures of the turn to data in ways that recognise their constitution in diverse social practices and specific situations but also how they circulate and get taken up for different ends. We did this by organising collaboratories across three different practical contexts: bioscience, national statistics and waste management.

We conducted a separate collaboratory for each of these three contexts, involving our team of social scientists and around 10 to 15 practitioners at every event. Each collaboratory was organised differently and variably comprised of presentations,

[6] Rabinow, P., G. E. Marcus, J. Faubion and T. Rees (2008). *Designs for an Anthropology of the Contemporary*. Durham, Duke University Press.

provocations and responses tailored specifically to their respective context. The idea was to explore the opportunities and challenges of working with Big Data by collaborating with practitioners who routinely use or are experimenting with data forms, and who share similar aspirations and apprehensions about the impacts of Big Data. Our project was, thus, both interested in collaboratories as a method of interdisciplinary and cross-sectoral engagement, and Big Data as a topic with a range of meanings and implications across practical settings.

Through these collaboratories we built on established connections and developed new relations for the social sciences with government and industry practitioners and experts not as end-users, but as collaborators who are part of the relations of production and interpretation of data. In this regard, despite our common methodological approach across the collaboratories, there were important differences in how we structured them. This was the result of our objective to trial different approaches; explore the different social lives of Big Data across and between practical contexts; and our interest in building on previous or ongoing working relations and/or establishing new relations with practitioners.

## Section 2: Three Context-specific Collaboratories

The following is a summary of the project team's initial framing and organisation of each of the three context-specific collaboratories.[7]

### Collaboratory 1: Genomics

Big Data came to genetics (and to biology) through the human genome project 1986-2001. The competitive nature of the HGP was a powerful impetus for the commercialisation and industrialisation of genome sequencing. Once completed, the human genome was translated from endpoint to starting point, and became the 'blueprint' for a new era of data-intensive science and medicine. This vision is being pursued, and since 2005 a new generation of instruments has dramatically increased the speed and decreased the cost of genome sequencing. In contrast to the fields of waste management and official statistics, genomics is now in a second phase of big data work that aims to leverage new biological and medical knowledge on the basis of vast pool of publicly available sequence data.

This First Collaboratory drew upon an established relationship between team members, Prof Adrian Mackenzie and Dr Ruth McNally, and UK genomic scientists. It built upon four years of research on genomic databases using a variety of methods (e.g. repurposing bioinformatics tools and scientific visualisations), some of which involved events and online encounters with UK genomic scientists. While those who participated in the collaboratory arrived with specific interests (e.g. research, commercial) and various forms of expertise – as a lab scientist, software producer or manager, for example – these prior relations informed the format of the collaboratory as part of an ongoing dialogue.

Genomic scientist speakers were invited to do presentations about their work in relation to the metrics for DNA and genomic data, whether as a lab scientist, software producer, a data user or a manager. The questions we asked them to consider included:

- What are the key metrics you rely on day to day? For longer term planning? For communicating with or persuading others?
- What can't you count or measure? What can't you count measure but yet still evaluate?

[7] Details of the proceedings of the collaboratories are summarised in a Supplementary Appendix, which can be requested by contacting one of the members of the Socialising Big Data project team.

- Are there things you once counted or measured, or were important to count or measure, but not any more?
- What new things are you trying to count or measure, or would like to count or measure if you could - and why?

Following their presentations the organisers provided a number of visual provocations that led to further discussion and debate. These included graphics and tables that made use of genomic researchers' own databases and software tools, and generally re-purposed them to raise questions about their metrics and ways of talking about the value of genomic sequence data.

### Collaboratory 2: Official Statistics

National statisticians have only recently started investigating Big Data as a potential data source for the generation of official statistics. Especially beginning in 2013, numerous international meetings and initiatives have been undertaken by Eurostat (the statistical office of the European Union), the European Statistical System (ESS, a partnership between Eurostat and National Statistical Institutes (NSIs) of member states) and the UNECE's Conference of European Statisticians. Additionally individual NSIs have been evaluating Big Data sources through, for example, the establishment of Innovation Labs to conduct experiments (e.g., Statistics Netherlands and the Office for National Statistics). Another initiative is the UNECE Big Data project 'sandbox' that provides a technical platform for NSIs to experiment with Big Data sets and tools. Examples include: analysing location data from mobile phones for generating 'real-time' population and tourism statistics; search query trends for generating data on migration; social media messages for generating indicators on issues such as consumer confidence; and price data on the Internet for producing inflation statistics. The sources and possible applications of Big Data are thus diverse including what is being measured and its relation to previous forms of measurement (e.g., surveys).

The Second Collaboratory on Official Statistics consisted of presentations by national statisticians from National Statistical Institutes (NSIs) and international organisations in Europe. Statisticians were requested to make brief presentations on Big Data related projects and initiatives within their organisations. An initial set of questions was provided to focus presentations about their current thinking about Big Data in relation to the question 'what counts?' and participants were also invited to pose their own questions in relation to this general theme:

- What can be counted or measured using Big Data sources? How are these different from or the same as existing sources?
- Does the use of Big Data sources call for new forms of statistical reasoning or tests or a 'paradigm shift'? How so?

- What can't be counted or measured using Big Data sources and why not? What is missing that you consider important?
- What would you like to count or measure using Big Data sources if you could and why?
- What does the use of Big Data sources for official statistics mean for the role of NSIs?

Presentations on their current state of thinking and experiments were then followed by questions and responses from the social scientists. The structure was in part a response to the fact that Big Data had only recently become an object of interest and experimentation among national statisticians. The presentations, thus, provided stocktaking of emerging approaches and understandings, building on recently established and ongoing working relationships between the practitioners and the organiser, Prof Evelyn Ruppert. Following the event, a paper was prepared outlining key themes and provocations that arose from the presentations and discussions. This was then distributed to the practitioners and responses were solicited in writing and/or through conversations at subsequent meetings and events. The collaboration subsequently extended beyond the initial event in an iterative process where the boundaries extended to a number of other engagements, interactions and conversations. This type of discursive exchange, as documented by the Anthropology of the Contemporary, enables practitioners to respond individually and collectively in the co-production of knowledge (with the co-production of collective work understood as an ongoing process). These iterations resulted in the reworking of the initial report.

### Collaboratory 3: Big Data and Urban Waste Management

Although there are many different types of data used in the waste management process, this is an area in development and the extent to which Big Data sources could be used to replace, supplement or verify existing data sources is as yet unclear. Waste management authorities are just beginning to investigate the possibilities of these sources.

The Third Collaboratory on Urban Waste Management was thus structured differently again, featuring 3 roundtable discussions involving a mathematician, social scientists, policy makers (Manchester, Birmingham) and waste management practitioners from UK local authorities (Manchester, Bolton, Stockport) under pressure to transform their services in an environment marked by austerity cuts and staff reductions. Similar to the other collaboratories, the roundtables were organised around a series of questions on the topic of Big Data reflecting the relatively new use of such data in the context of waste management and in relation to methodological issues, ethics, openness, policy and behavioural change.

- How does big data differ from other types of data? Are new measures produced? How does data differ from information? How does big data do counting and measuring differently to statistical or administrative data? What is measured and what is valued?
- What are the possibilities and challenges of working with big data in urban waste management? Is big data open data? What are the possibilities and challenges of public-private partnerships for data management?
- How can big data be used to shape policy decisions and respond to future challenges in waste management? Does it allow a different relation to the public?

Unlike the other collaboratories, however, it was based on a recently established relationship between the organisers, Prof Penny Harvey and Dr Camilla Lewis, and UK practitioners who were in some cases, already familiar with each other, and between Prof Celia Lury and Birmingham City Council. It also involved the co-creation of the collaboratory content and format. Additionally, the interweaving of academic and practitioner presentations brought to the fore different ideas and understandings of the issues and concerns at stake in working with Big Data.

So while our three collaboratories adhered to a common collaborative method in their commitment to experimentation, discussion and debate, the approach to each practical context varied according to the specific contexts and practices and our relations to them. Moreover, while we continue to maintain that collaborative endeavours of this kind require a commitment to openness and uncertainty by relinquishing preconceived truths or definitions, we recognise that epistemic asymmetries exist both among and between participants. This had an impact on the capacity of all collaborators to participate and contribute equally to a 'shared literacy'. These power dynamics and extant inequalities in terms of skills and expertise, especially in relation to technocratic issues of working with Big Data, were some of the political and practical challenges of our methodological approach.

Following from our three collaboratories, which formed the initial part of our multi-method approach, we conducted two postgraduate workshops in June 2014: one at the ITU in Copenhagen and another at the London Social Science Doctoral Training Centre. Together, the three collaboratories and postgraduate workshops informed the organisation of a final collaboratory.

## Section 3: Final Collaboratory – Discussion of Findings

This final collaboratory brought together participants from the three context-specific collaboratories. The format was chosen as a way to consolidate and reflect on the findings of the first three collaboratories and for practitioners to learn from experiences and concerns in relation to contexts that have different histories and trajectories of working with Big Data. In brief, genomic scientists are entering a middle phase, official statisticians are beginning an experimental phase and waste managers are initiating an exploratory phase. The practitioners also occupy different positions in relation to the analysis and application of Big Data, from policy and service provision to statistics generation and scientific research. Despite the fact that genomic scientists, statisticians and waste management practitioners approached Big Data for different purposes and from different perspectives, the collaboratories enabled the Socializing Big Data team to identify affinities in how it was understood. These were described in advance of the final collaboratory as 'crosscutting themes' - metrics, economies, ethics and collaboratories.

The collaboratory also involved discussion of the initial formulation of 'Socialising Big Data.' We had started with the social scientific assumption that this formulation would be of concern and interest to practitioners working in different contexts. By bringing attention to the 'social lives' of Big Data, our objective was to explore the social and technical practices that generate, organise, curate, circulate, and analyse Big Data, highlighting that these are not neutral but consequential to what is known. Furthermore, because the social lives of Big Data are becoming ever more distributed and dispersed, these consequences are difficult to ascertain. Our format acknowledged that the meaning of these issues, and the way in which they play out, are specific to different contexts. This was why we considered it imperative to engage with three different practitioner groups. While for us this approach led to many insights and understandings of the social lives of Big Data, the extent to which this understanding is and can be meaningful to practitioners was a question that we also posed for the final collaboratory.

The following sections consist of the project team's analysis of the first three collaboratories in relation to four crosscutting themes that arose in each context and which were presented and discussed at the final collaboratory along with questions and possible policy implications. Together these in essence constitute the team's analysis of the findings from the first three collaboratories.

## Crosscutting Theme 1: Metrics

### Context: Genomics

Genomics practitioners inhabit a diverse data ecosystem in which they undertake differentiated yet interdependent roles. These roles can be categorised into 3 ideal types: gleaners and cleaners; packers and stackers; and action heroes. Each role entails different metrics and only partially overlapping metrics:

- Gleaners and cleaners: People working closely with sequencers take a strong interest in speed and cost of sequencing. Their metrics include cost/genome, hours/genome. In the last few years (2011-), a target price of $USD1000 has been constantly discussed. Metrics relating to the production of sequence data also relate to the reliability and accuracy of sequence data. Producing genome sequence data is not a simple capture and recording operation, but involves many processes of collecting and preparing samples, or assembling sequence fragments into a composite whole. Metrics relating to this process are commonly discussed by practitioners in talk about coverage, read-depth, etc.
- Packers and stackers: People working mainly with sequence databases use metrics relating to data volumes and data traffic. They are keenly interested in metrics concerning data compression, data transfer speeds, and discoverability. Genomics data moves between commercial and public data platforms, and metrics comparing different platforms such as Cloud compute have been widely discussed. Costs of moving, copying and processing sequence data are often discussed, and lead to metrics such as 'doubling time' that allow practitioners to plan storage or computing needs. Metrics relating to data quality, consistency, and quality of associated metadata.
- Action heroes: People making use of sequence data to understand biological function use an entirely different set of metrics drawn from diverse domains of biology and medicine. These metrics are often much more statistical in character, and largely concern differences between sequences. They extensively measure similarities and variations between closely related sequences in order to, for instance, calculate risk or biological relatedness. They make increasingly heavy use of predictive models, so metrics relating to error rates, sensitivity, specificity, etc. are common.

*Questions*

- How do the practitioners from the other two contexts (waste and official statistics) map onto the 3-role ecosystem? Are they only one type?
- Do different practitioners in these other domains relate to different kinds of metrics? If so how and why?

- What metrics are missing? While metrics for speed and cost abound, where are the metrics for realisability of promised advances, or even metrics for results of the last two decades of genomic research?

*Policy implications*

- Need to develop wider variety of metrics for genomic data, including metrics of data reuse, metrics of data linkage, etc.

**Context: National Statistics**

It is often noted that official statistics are based on a 'design and then collect' process whereas Big Data is based on a 'first collect and then design' process. While sometimes referred to as data-driven or technology-driven inquiry, a 'design follows collection' approach demands that the various practices involved in 'designing' Big Data need to be transparent and evaluated. This requires access to how data is generated, processed, cleaned, organised and provided (e.g., how search engine queries are ranked) and understanding the implications for different kinds of questions and analyses.

At the same time, platforms such as browsers and social media are unstable and changeable, which raises questions about data quality, and make metrics and measures unreliable and longitudinal analyses problematic. For example, studies of Google Flu Trends illustrated how search queries can become less reliable over time due to the changing behaviour of users. In general, many Big Data sources are measures of behaviour, of what people do, including their patterns of opting in and out of platforms, creating multiple online identities, and inconsistent or irregular use of platforms. These issues potentially make Big Data sources incomparable and meaningful only in relation to specific platforms, moments or issues. While agreements with platform owners – either through PPPs or specific-use arrangements – can possibly address these issues, because platforms are not designed for 'statistical purposes' qualifications in the use of these sources are required.

In the face of uncertainty and questions about quality, instead of generating measures, Big Data can be complementary to official statistics such as providing new measures (e.g., ICT usage, tourist movements) and supplementing/verifying existing ones (e.g., sentiment analysis in relation to surveys). Rather than appealing to standard statistical measures (e.g., averages) or tests of validity, Big Data can also be used in unique and more 'timely' ways such as providing 'first warnings' or 'signals' through the analysis of patterns (e.g., search queries indicating emerging issues) and trends (e.g., mobile phone data indicating changing movements). Based on these, more in-depth investigations concerning questions such as causality can then be

undertaken (such as the approach followed by the UN Global Pulse). Visualisations of trends, for example, offer a powerful way of displaying data to compare and identify correlations and possibly causation. Through such a 'responsive' mode NSIs can develop the capacity to analyse and interpret Big Data and introduce innovative ways of understanding changing societal practices and processes including the ways they are being rendered into data.

Because it is relatively novel in official statistics, working and experimenting with Big Data is necessary to test its qualities, uncertainties, capacities, and so on. It involves fluid and serendipitous processes. In this regard, Big Data can be the basis of experimental projects using modelling and simulation techniques that can provide a space for identifying both problems and possibly solutions but also in ways that can be complementary to official statistics. However, these approaches rely on hunches, guesses, intuition and speculative searches, which are not independent of hypotheses, theories, assumptions, and pre-conceived notions. This is especially evident in interpreting and differentiating between 'signals' and 'noise' or 'babble'.

*Questions*

- In a time of decreasing resources, how can Big Data experiments be justified and promoted?
- What are the organisational barriers to working with Big Data sources and the different analytics and understandings of evidence that they call for?
- What are the benefits and challenges of international collaborative initiatives for experimenting with Big Data sources?

*Policy implications*

- The pressures of responding to existing user/stakeholder demands means that exploratory and experimental work is difficult to justify.
- Working with indicators, signals, trends and patterns introduce speculation and uncertainty, which demand careful explanation and interpretation.

**Context: Waste**

Although there are many different types of data used in the waste management process, this is an area in development and the extent to which big data sources could be used to replace, supplement or verify existing data sources is as yet unclear. All waste collection authorities must report to the national monitor Defra, via the Waste Data Flow system which records the tonnages of waste which are collected. There is therefore, a lot of data at the national level but less is known about data at the household level. In some countries, e.g. in Scandinavia and in Spain, sensors technologies are used to provide real-time information on what wastes are being disposed of and by which households. This information can be used to move away

from fixed waste collection rounds towards more responsive systems that calculate optimum transport routes according to the materials that are in the bin. These technologies could also potentially enable waste authorities to pinpoint households that fail to properly sort waste for recycling and reuse. The data could thus be used to create tailored services that are anticipatory and predictive. It could open new possibilities of modelling, simulation and forecasting. However, critics argue that it is not clear that such data would create new knowledge beyond more timely and precise information about the number of bins collected. It is therefore important to clarify how the data generated from these devices would qualitatively change the information available to the authority.

*Questions:*

- What is the transformative potential of Big Data for waste management?
- Would chips in bins produce new metrics, or would they replicate existing metrics?
- Could real-time data be used alongside existing data metrics to facilitate a better policy debate?
- Could big data be used to help waste authorities meet new EU targets?

*Policy implications*

In July 2014, the EU will introduce a new Waste and Circular Economy Package that is expected to change the way in which data on waste is collected and used. The EU will require all member states to use a single data methodology to define the success of recycling rates and to standardise information on the quality of materials being recovered and circulated in the economy. The new form of measurement will require waste disposal authorities to record the quantity of materials recovered for recycling. The goal is to have 50% of municipal waste being recycled by 2020. Recycled materials are increasingly treated as commodities and traded in global commodity markets. Understanding waste as a commodity raises new concerns and possibilities in the face of a growing awareness of finite resources, the need to protect the natural environment and improve responses to resource scarcity.

## Crosscutting Theme 2: Economies

### Context: National Statistics

Cost savings from reusing existing Big Data sources is a key benefit but at the same time Big Data introduces new costs, some which are difficult to know and evaluate. In addition to requiring investments in IT, training and hiring staff, and developing new methods, there is uncertainty about the costs of using and possibly having to purchase data from commercial owners in the immediate or long-term. Cost considerations are therefore both a driver but also a source of economic uncertainty and vulnerability for NSIs especially at a time of budget constraints.

There is as a result much interest in building public private partnerships (PPPs) with commercial data owners towards securing access to and potentially reducing and/or fixing the costs of data. This is also desired on a cross-border and international basis since the data generated by major platforms (e.g., Google, Twitter) transcend national boundaries. For some statisticians, PPPs are understood as a necessity as commercial owners are 'ahead of the game', investing more in Big Data than NSIs and attracting the best talent. While at one time the statistics provided by NSIs were unique, other players have entered the information market and have started generating statistics, for instance, on inflation and price indices. NSI initiatives also need to be understood within the broader context of EU-wide plans to develop capacity and share data, infrastructures, skills and legal frameworks in the building of a digital economy.

The valuation of Big Data as a source for generating official statistics is also a result of pragmatic considerations. If Big Data can provide answers to questions that matter and do this in a more timely fashion than standard methods, then policymakers and other users will be better served. If official statistics can't answer questions that matter then they will not be relevant or valuable. Investing in Big Data may thus be worth the costs even if they are uncertain or higher.

NSIs can show responsible statistical leadership through advancing the UN Fundamental Principles of Statistics (impartiality, reliability, relevancy, profitability, confidentiality and transparency) in relation to Big Data sources. This could include providing accreditation or certification on different data and measures. In this way, NSIs could contribute their experience and skills working with, validating and linking diverse data sources and generating a wide variety of statistical outputs. This is one possible role for NSIs – as trusted third parties - in the Big Data valuation chain.

Generally this suggests a changing role for national statisticians, from producers of bespoke data to analysts of data produced by others and for other purposes or what is suggested in relation to the distribution of roles in genomics as a change from

'gleaners' to 'action heroes.' Whatever the designation, Big Data raises the question of the distributed relations involved in the economies of Big Data, including those distributions within NSIs, where data science is understood not as requiring the skills of a particular person but distributed amongst a team including methodologists, statisticians and IT people.

*Questions*

- What are the implications of national or transnational PPPs in the valuation and legitimating of Big Data as a source of 'official' statistics and their role in the formation of a 'data driven economy'? What are the costs and benefits?
- What does using Big Data mean for the 'independence' of NSI's in the provision of 'high quality information?'
- Does the distribution of Big Data skills, ownership, technology and innovative analytics in the private sector put national statisticians in a defensive position?

*Policy implications*

- There are upfront and long-term costs of developing Big Data applications and methods and these have implications for the resourcing of other NSI activities.
- Big Data development work needs to link and connect to the policy interests and needs of government departments and those of different stakeholders and users of national statistics.

**Context: Genomics**

**Description:** Big Data came to genetics (and to biology) through the human genome project 1986-2001. The competitive nature of the HGP was a powerful impetus for the commercialisation and industrialisation of genome sequencing. Once completed, the human genome was translated from endpoint to starting point, and became the 'blueprint' for a new era of data-intensive science and medicine. This vision is being pursued, and since 2005 a new generation of instruments has dramatically increased the speed and decreased the cost of genome sequencing. In contrast to waste and official statistics, genomics is now in its second phase of big data work that aims to leverage new biological and medical knowledge on the basis of vast pool of publicly available sequence data.

At the level of production of data, the production of sequence data by next generation sequencing machines, and hence the volume of data flowing into databases is closely associated with market competition between the major manufacturers of sequencing machines (Illumina, Pacific Biosciences, etc.). These machines in turn are shaped by the different investments in genomic research. The economies of genomics focus around biomedical applications and are arguably increasingly dominated by large sequencing centres such as the Sanger Centre (UK),

the Beijing Genomics Institute (China) and Broad Institute (USA). Much of the scaling up of genomics from single individuals to large cohorts of people seeks to address the problems of finding variants associated with disease or propensity to disease. Hence, the genomics research landscape is dominated by large population level consortia projects that produce huge amounts of sequence data, are often highly international and involve hundreds of researchers. At the same time, genomics research, like many data-driven enterprises, has been heavily committed to personalized medicine. The promise of individual whole genome sequencing as well as the popular of individual genotype profiling (as marked by 23andme) has led to desktop sequencing instruments, to a proliferation of genome-wide association studies on a wide variety of medical conditions, and above all to a much intensified focus on translating genomic research into clinical settings. Nearly all of these developments rely on the public availability of most of the sequence data in public databases. DNA sequence data functions almost as a public good in the sequence data economy.

A second distinctive economy associated with genomic data concerns the remit of genomic data. Sequence data has gradually become ubiquitous in many different life sciences, ranging across medicine, drugs, health, agriculture, biotechnology, renewable energy, environment and many other fields. As applications of sequencing have broadened, uses and techniques of analysing sequence data have expanded, but often in tension with existing scientific expertise (for instance, plant breeding vs. genetic modification; ecological field study vs sequence-based studies).

Genomics has long had its own version of the 'big data' skills shortage. Beginning with the Human Genome Project in the early 1990s, the 'bottleneck' in genomics (that is, its difficulty in delivering on the promise of deep biological understanding through sequencing) has been attributed to shortages of people able to analyse the data. Whether this skill shortage has disappeared or not, analysis of genomic data is still seen as the most expensive and time-consuming part of genomics. It has been addressed by changing infrastructures (for example, the increasing use of Amazon Web Services or Google Compute), through the growth of commercial sequence data management services, and by sequence machine manufacturers themselves in the development of algorithms and software. Needless to say, algorithms and techniques developed initially for bioinformatics and genomic research have filtered out into other data-intensive sciences.

*Questions*

- Does the mixture of public and commercial interests in genomics data offer any guidance for other big data settings?
- How does the long-standing skills-bottleneck in genomics suggest what could happen in other domains?

*Policy implications*

- Need to ensure that a single form of data practice does not homogenise or dominate domains to the exclusion of other ways of acting, knowing or relating.
- Need to invest in forms of skills training that are not too focused on current problems or technical difficulties but countenance ongoing change?

**Context: Waste**

**Description:** Local authorities are under increased pressure to transform their services but they are also faced with austerity cuts and staff reductions. Public services, such as waste management, are increasingly managed in public/private partnerships in which new markets for specialist service provision have become central. Data is central to these partnerships both in terms of targets and agreements, and as a valued resource in their own right. From the perspective of local authorities, new data solutions might require considerable financial investment and political will. However it is difficult to be certain that such investments will pay off as the potential of big data outcomes lies primarily in the uncertain possibility of generating new, unexpected perspectives. Big data could also be used to offer financial incentives to users. There have been a number of trials where a rewards system (similar to supermarket loyalty cards) has been introduced to incentivise recycling. With this technology, it is possible for individuals to 'opt in' and trace their personal information and thereby support an ethic of participation. However, there is a high level of mistrust about putting sensors into bins and concerns that individuals' information could be misused, especially now that big data analytics contribute to making data a commodity in contexts where it is not always easy to ascertain who reaps the benefits and how.

*Questions*

- What are the economic benefits of engaging with the private sector on waste reduction?
- What are the commercial advantages of engaging with the private sector with big data?
- What kinds of data would the private sector be interested in gaining access to from the public sector?
- How could the public sector engage with the private sector on Big Data questions?

*Policy implications*

- A tension is apparent between the unknown potential of big data and the requirement for waste practitioners to produce results (i.e. to make specific things happen) within relatively short time frames. There are a number of

different private partners who it may be fruitful to engage with such as construction companies and manufacturers who produce food packaging. These companies also have a social responsibility to reduce the quantities of waste they produce and could work alongside policy makers to tackle these issues.

### Crosscutting theme 3: Ethics

**Context: Waste**

Contemporary waste management relies on specific data flows particularly of the tonnage data which tracks both the total weight of waste collected, and the proportion of waste that is processed at recycling plants. This tonnage data is used to map habits, project trends and calculate recycling rates. Despite the volume of data generated by waste management authorities in the UK there is as yet little systematic engagement with Big Data metrics, although the waste sector is beginning to experiment with such data forms, particularly through the use of 'smart bins', equipped with sensors. Such sensors could generate data with the potential to provide new kinds of information helping waste authorities to identify patterns and make predictions based on large quantities of real-time, detailed information. For example, Big Data could provide valuable insights into the possible correlations between 'accurate recycling' and other variables (e.g., responsive collections, the use of incentives and disincentives, weather patterns, etc). The combination of the volume of data and its timeliness suggest that correlations, as yet not fully imagined, could also emerge and be tested. However the re-purposing and linking of data sets raises ethical questions about informed consent. There are also issues of trust associated with the risks of false or spurious correlations and fears about invasion of privacy, that need to be addressed.

*Questions*

- How might new ways of measuring data, such as sensors, introduce new concerns about the relationship between the public bodies and private concerns? Should limits should be placed on big data analytics? How might decisions about such limits be openly discussed and debated?
- Could chips or sensors on bins support the idea of ownership of bins, encouraging individuals to think and behave in new ways with respect to the disposal of domestic waste?
- Could Big Data to be used to target interventions and what would be the implications of this approach be in the wider field of public service delivery?

*Policy implications*

- Senior Officers of the GMWDA would like to see an informed policy debate on the potential of installing chips or sensors in bins as the data generated could enable them to provide a more efficient and personalised service. However, they are cautious as there have been a number of campaigns in the media which stress that chips or sensors in bins might also have negative consequences, raising fears about surveillance technologies, that allow and even encourage the authority to 'spy' on households.

**Context: National Statistics**

Ensuring the privacy and confidentiality - both actual and perceived - of personal data are key concerns of NSIs. On the one hand, a variety of techniques are used to achieve this such as anonymisation, disclosure protections and the transparency of practices. On the other, there is an assumption that reusing existing data is less intrusive and demanding of respondents and thus more respectful of privacy. However, even when data is anonymised and thereby no longer 'personal', the repurposing and linkage of different datasets may lead to the identification of individuals as well as 'group effects,' where the identification of patterns and relationships can be used to target particular groups resulting in further concerns about the reuse of data.

Big Data sources such as social media and mobile phone usage also raise the issue of consent. NSIs have longstanding practices of making transparent to respondents the intended uses of their data and any changes are subject to stringent data protection review and approval processes. How this can be accomplished in relation to Big Data (from social media or search engines, for example) is currently a matter of review and debate. For example, data protection rules generally stipulate that consent must be freely given, specific and informed. However, what users of specific platforms originally agreed to may not cover third party use and repurposing of data. Furthermore, the criteria that consent be specific and informed means subjects must be aware of the purposes to which their data may be put ('purpose limitation') and new purposes must be 'compatible' with those stated purposes. Though there is much debate about what constitutes compatible uses, this criterion potentially conflicts with the exploratory and serendipitous character of Big Data experiments where uses are 'discovered' in the data (as discussed under metrics).

In addition to or in place of identifying policies and procedures that can address these issues, two other approaches are possible. One involves using Big Data for measuring things (crops, water, prices, traffic) rather than the doings of people.

A second concerns using data collection process data (paradata) to improve existing methods. With the increasing move to online censuses, government services and surveys, a large amount of paradata (usage and behaviour data such as clicks, duration, pages read, and field operational data) is being generated and which could be used to experiment with 'in-house' forms of Big Data. While still raising questions of ethics, if made transparent by NSIs, paradata could be used to improve data collection processes. This would be more 'low risk' and enable early experimentation and capacity-building in working with new forms of data as well as contributing to changing organisational cultures.

Big Data sources generated by commercially owned platforms are also vulnerable to privacy and ethical controversies that publicly erupt as a consequence of revelations about surveillance, tracking and data sharing. By using these sources, NSIs also become vulnerable and possibly implicated in these controversies.

*Questions*

- What are the ethical risks of using Big Data and how and to what extent could they be addressed?
- Might the data protection and privacy approaches of national statistics and governments more generally be their 'competitive advantage' and serve as a basis for the development of approaches in the private sector?

*Policy implications*

- While data protection principles are well advanced in relation to government data, they are not so for Big Data sources. Initiatives currently underway such as the proposed EU General Data Protection Regulation, may provide policy guidance on this.

**Context: Genomics**

A minor academic and professional industry has developed around the ethical, social and legal implications of genomics as data-intensive science. In Europe, UK, North America and several other countries, government-funded research has extensively researched ethical issues associated with genomics, mainly in the interests of protecting patients, citizens and public in general from either losing control of their own data, or in the interests of helping various social groups manage potential disadvantages or discrimination associated with genetic data.

Explicit ethical issues around sequencing data are legion, and include generic 'big data' concerns such as personalization and de-anonymization. A recent study showed for instance that it was possible to identify named individuals from genome sequences deposited in public databases. While the international genomics community has carefully architected databases to guard the confidentiality of clinical

sequence data (for instance, maintaining separated access-controlled database such as dbGAP), public sequence datasets can be de-anonymised using relatively straightforward data linkage techniques.

This problem is complicated by the increasingly commercial-hybrid character of much genomic research. Large sequencing centres effectively operate as global sequencing services for clients. Cloud computing services such as Amazon Web Services and Google Compute are not subject to the same regulation as publically funded research. Use of these platforms for sequence data is troubled by issues of trust, and uncertainty as to the commitment of service providers to the ethical frameworks that bind genomics researchers.

We would suggest that many framings of ethical issues associated with genomic data have been narrowly individualistic, and they have paid little attention to ethics already implicit to data practices. As it moves between different settings -- research setting, clinical research, clinical application -- biomedical sequence data is valued differently. Error margins or acceptable risk differ between a research laboratories, industry research and clinical settings. What seems highly promising to a laboratory-based genomics researcher might be highly problematic to a clinical practitioner or public health professional.

*Questions*

- How deeply are ethical concerns carried into data practice?
- In what ways does an ethic of care already operate in data curation?

*Policy implications*

- Investigate and encourage a wide range of involvements in setting agendas and priorities for genomic research
- Do not assume that ethics only relates to human research subjects or patients but also plays out in many different forms of relationship.

### Crosscutting theme 4: Collaboratory

**Context: National Statistics**

National statisticians currently work collaboratively via numerous forums such as those facilitated by Eurostat or the UNECE. Generally, these collaborations involve statisticians who are similarly positioned within NSIs in either management or project roles. While collaboration is thus not new to national statisticians, methods for doing this with social scientists and the private sector - beyond meetings and stakeholder consultations - are not as well established. Big Data provides an opportunity for developing such cross-sectoral collaborations for many reasons. For one, Big Data is ‘new’ and there is little settlement on applications, methods and

consequences and thus more openness and experimentation rather than already settled positions about its possibilities. But perhaps more significantly, because of the very nature of its production and potential applications, many different interests and players intersect with Big Data, which necessitates investigating methods of collaborating. And finally, amongst NSIs new forms of collaboration are being experimented with such as the UNECE sandbox project. Rather than individual NSI's developing new methods and then sharing these with others as best practice (which is the usual process), the sandbox is intended to involve collaborative experiments with Big Data.

Because of the relative newness of Big Data as a potential source for generating official statistics the collaboratory involved a few iterations (detailed in the Supplementary Appendix): a workshop-type event involving 'stocktaking' discussions of Big Data related projects within NSIs in response to some initial questions; email distribution of a follow up summary and analysis of the discussions; and further documentation of responses to the summary and analysis via subsequent individual conversations and meetings. This process thus involved on-going conversations about the issues raised at the initial event.

The collaboratory was thus not organised to share skills, develop methods or analyse Big Data, but to pose critical questions (e.g., what can and can't be measured, what is valued) about its methodological and political implications for official statistics. This reflected the aspiration to bring into conversation the different interests of social scientists and statisticians, such as epistemological questions of method in relation to practical demands for the production of relevant official statistics. For social scientists, the discussions about and understandings of Big Data in relation to official statistics usefully inform their research and teaching. The benefits to practitioners involved on the one hand building relations with social scientists and posing and addressing questions about Big Data they might not otherwise. On the other, it was a different context for practitioners to meet with each other and share experiences. But given the framing was lead by the social scientists – that of 'socialising Big Data' - how this framing was interpreted and whether it was meaningful to statisticians and benefited their practical work was a question opened up at the final collaboratory.

*Questions*

- What formats would be possible to enable participants to switch roles, that is, for collaboratories to be multidirectional in setting agendas and issues (i.e., between social scientists and statisticians)?
- How can collaboratories be left open to 'productive misunderstandings'? That is, rather than seeking consensus, how might we state the values or benefits of

bringing together and allowing for tensions and different perspectives and interests to be expressed and engaged?
- What are the benefits – both experienced and desired – of collaboratories on the part of statisticians?

*Policy Implications*

- In the face of economic and practical constraints such as time, the relevance and value of collaboratories that are more exploratory and conceptual rather than directly instrumental (e.g., developing specific applications/methods) need to be outlined rather than assumed.

**Context: Genomics**

Genomics research sprawls across industry, education, government, and business. It is widely distributed, and increasingly carried on at many different scales ranging from citizen science to global consortia, from lab or desktop sequencing, to massive population-level studies. The variety of fields and settings intersected by genomics and sequencing techniques can make it hard to identify coherent problem domains or debates. The long-standing promise of sequence data as a digital readout for biology, and the long-established ethical and legal discussions around genomic data can make it difficult to establish collaborative relations with genomic researchers. There are simply too many different interests, voices, and initiative going on in genomics to bring to one table.

Many genomic researchers are well-versed in the main ethical and social issues associated with genomics. In biomedical settings, researchers and practitioners are highly sensitive to ethical issues, especially because ethical reviews are part and parcel of their research planning. In some cases, genomic researchers have been repeatedly interviewed by social scientists and even mainstream media, and these experiences inform their approaches to any dialogue concerning genomic data. This familiarity with ethical and legal discussions can make it difficult to initiate other topics of dialogue.

For instance, how does one start discussions around genomic economies or metrics? We found it necessary to put discussions on a different footing by working in visual terms (graphics, tables), making use of genomic researchers own databases and software tools, and generally trying to re-purpose genomic researchers own data literacy in the conversation by showing them data gathered from databases about their own data. This approach leads to mixed results. On the one hand, it certainly overcomes some problems of distance and unfamiliarity. That is, the genomic researchers are looking at the kind of data that members of their own community might use. On the other hand, this data is now presented with a view to challenging

them to think about their own metrics and their own ways of talking about the value of sequence data. Some robust discussion usually arises.

Notable differences in collaboration can be seen in different areas of genomics. We found clinical researchers difficult to engage. They have little time. By contrast, more junior and post-doctoral level researchers are often quite curious and interested.

*Questions*

- At what places and times are conversations about 'big data' likely to be most engaging?
- Is it relevant to consider collaborative work that varies according to the experience of the practitioners?
- Could one envisage multi-sited collaboratories? In certain complex and vast 'big data' domains, this might be useful.

**Context: Waste**

The prime responsibility of a waste authority is to deliver a service to the public, and officers have concerns around committing themselves to exploring the uncertain potential of Big Data. Academics on the other hand are expected to explore such possibilities in more open-ended ways and may well be able to provide authorities with new questions and/or perspectives to stimulate debate rather than simply offering 'solutions' to pre-defined 'problems'. In this way, collaborations could make it possible for local authorities to engage more experimental approaches without diverting core resources. One possibility is that 'urban laboratories' could be set up as partnerships between waste management authorities and social scientists for the design and conduct of experiments in evidence-based research. These laboratories would be collaborative spaces in which public bodies and academics would negotiate the tensions between the need to 'make things happen' and the potential, but uncertain, benefits of exploring possibilities in an open-ended way. For example, research shows that there are strong links between infrastructural variables and recycling rates. Recycling rates in highrise flats and in areas with a high turnover of people tend to be much lower than in neighbourhoods where individuals have easy access to recycling bins and reliable collection services. Equally it is common knowledge that there are key moments where individuals throw away large quantities of waste, such as moving house or after the death of a relative, but the 'lumpiness' which is caused by these incidents is usually written out of large data models. Big Data analytics generate correlations and patterns that could be empirically tested by academic researchers. Thus for example if the data shows that recycling rates are lower in apartment buildings the researchers could test and compare variables – without assuming specific lines of causality. Behavioural and infrastructural variables could be looked at together.

*Questions*

- What are the benefits of engaging with academics on Big Data questions?
- Is the collaboratory format a useful way of furthering such engagements?

*Policy implications*

- Devising and carrying out experiments could become a fruitful site for on-going collaboration between waste practitioners and academics and could inform policy makers.

## Section 4: Summary of Key Conclusions

The crosscutting themes, questions and policy implications generated much discussion and debate and some of the key points were well captured in the concluding session, which is summarised below.

In evaluating the collaborative approach, Celia Lury reflected on the different pronunciations of the term, 'collaboratory', which she suggested reflects something of the history of the term. Whereas the term 'co-laboratory' emerged in a scientific context, the 'collaboratory' as a collaborative method is more widely used in the humanities and social sciences. Celia contended that both inflections are useful. Employing the collaborative approach in relation to Big Data has shaped how the team organised the collaboratories in terms of the kinds of questions asked and the practitioners with whom we collaborated. Interdisciplinarity is often precipitated by a notion of crisis, the idea that there are pressing problems that require disciplines to come together. Big Data is an emerging field that disrupts and challenges standard working practices and lends itself to interdisciplinarity and asking questions such as: What is Big Data as a problem space and how can we this space through different modes of collaboration? Big Data involves a redistribution of data collection and research methods expertise and the restructuring of infrastructures, which necessitate engagements with a wider range of collaborators. In order to address questions around the social life of Big Data then requires engagement with practitioners from both the public and private sector.

From a social science perspective, collaboratories can provide a testing ground for concept development. It is important to consider whether we have learnt anything about the kind of 'socialising' involved. For example, what are the frameworks for thinking about Big Data? In terms of policy, we have legal, economic and political frameworks for thinking about Big Data. Should we add a social framework for thinking about Big Data and, if so, how would a social framing be different from these existing modes of analysis? From this perspective, collaboration may be thought of as an iterative process distributed not only in terms of space, but time. In terms of knowledge production, collaboratories bring social scientists into the collaborative process from the outset rather than merely being there to challenge, critique and problematise the findings of social scientific research. What is exciting about collaboratories is that they help us to move beyond individualised disciplines and projects by providing a method to develop and tests concepts.

Hannah Knox, from the Dept, of Anthropology, UCL responded to the discussions by reflecting on the genesis of the project. She noted that the collaboratories were conceived at CRESC as a way to make academic research more useful and to have a

greater impact. The method was designed as an experiment to trial the impact of opening up communication by assembling people (researchers, stakeholders and practitioners) at the initial stage of questioning and agenda setting rather than merely documenting findings towards the end of a project, as is typically the case in academic research.

In light of these objectives, Hannah emphasised the interrelationship between collaboratory as method and the topic of Big Data. When problematised, Big Data requires particular forms of collaboration between different stakeholders and practitioners. Despite the will and ambition for collaboration, commercial and political interests can act as powerful boundaries to collaboration on the topic of Big Data. This Final Collaboratory provided a neutral space in which to discuss some of these challenges, such as attempts to integrate Nectar card data from Sainsbury's loyalty card schemes with that of other organisations, which was blocked due to Sainsbury's existing relationships with other commercial enterprises. In this regard, collaboration provides a useful way to understand the problems of working with Big Data. Through this approach, for example, we can identify who the important players are and ask questions about this burgeoning topic. The Final Collaboratory has revealed some of the key players in the field, but certain stakeholders were absent, such as, the users and producers of Big Data.

Hannah ended her presentation by thinking about how to proceed with the collaborative approach. She emphasised the value of developing a shared vocabulary, but was curious about whether this would take an oral or written form (via publications or an extension of the working paper, for example). She then asked whether collaboratories would lead to new modes of experimentation or novel research projects, concluding by highlighting the importance of talking collectively about the benefits of collaboration as a method.

The group then engaged in a general discussion and raised the following points about the collaboratories and what was accomplished.

- What has been started here should not sit on a shelf; this was just a beginning.
- One of the outcomes has been the establishment of a diverse network of people engaged in questions of Big Data. Out of this we could consider possibilities such as a project involving waste management authorities, ONS and social scientists.
- The project has widened horizons and enabled connections that might not otherwise have happened. The diverse and conversational approach of the final collaboratory was appreciated; it enabled people to speak without the fetters of 'credentials' and provided a safe environment to think out loud. That said, more provocation and controversy could have been introduced.

- The working paper was especially helpful. But an alternative approach to the structure of the collaboratories would be good to consider. The position of the social scientists seemed to be more as observers rather than active participants. It would be good to consider a model that is more of a mix.
- The insights of the project need to come forward especially in the face of documents such as the EC data driven economy – why not think about a data driven society?
- More private sector involvement would be a good next step as well as from data scientists, privacy groups, data journalists and so on. Additional follow-up actions would be good to identify.
- It is good to talk about Big Data but what is also needed is a space for not just flying ideas but doing Big Data that could support the move to policy development.
- The concept of socialising is useful for understanding the different norms of different disciplines and interests and the benefits of mixing or 'socialising' them.
- How might the international aspects of Big Data be better leveraged? Recognizing that Big Data generated by online platforms cuts across national borders it would be useful to have forums that address this.

In a further, iterative response, the project team decided to build on the proposal to develop a 'social framework' for the use of Big Data, which will be the subject of a further publication. This involves a return to our original term 'socializing' Big Data, which we now believe has at least two senses.

The first sense relates to our original hypothesis, and which the collaboratories confirmed, is that Big Data is not a simple or unitary category, but *has multiple histories and contexts of use*, *which are being folded into the formation of Big Data itself*. One immediate proposal here is that Big Data be recognised as a plural or collective, rather than singular, noun, as a way for emphasizing that it is not a unified or consistent whole. In short, Big Data is not a fixed entity, but is in the process of being composed, and as such involves not only data-sets, but practices of collection, techniques of analysis, methods of storage, and relations with users, etc. It is an emergent socio-technical assemblage – perhaps best described as a Big Data multiple, in which current practices, across a range of fields, have the potential to profoundly influence what it becomes. Given the multiplicity of Big Data, a further proposal towards the development of a social framework is that there is a need for all actors to participate in the creation of a shared literacy or Big Data lexicon. This will require an understanding of how the diverse histories and contexts of use are shaping Big Data.

The second sense of socializing Big Data that we wish to draw attention to here is *its capacity to socialize*. As one of our collaborators put it in the final collaboratory, Big Data is inherently social, that is, its still as yet undefined potential is tied to its capacity to establish relations within and outside itself - to multiply, to divide, to provoke the creation of new data, to replace other ways of knowing and to provide the basis for new kinds of evidence, informing the activities and decisions of government, business and individuals. What we think is at stake here is understanding how relations between data are also simultaneously relations between people. This is not to say that there is a direct or one-to-one mapping here, but to acknowledge that data is never simply closed or already formed.[8] As Whitehead writes of number, 'The very notion of number refers to *the process* from the individual units to the compound group. The final number belongs to no one of the units; it characterizes the way in which the group unity has been attained' (1968: 93).[9] This is a moment in the emergence of Big Data similar to that in the twentieth century in which the state's policies came to be directed through the construct of the 'statistical personage'. Focusing on the capacity of Big Data to socialize would enable us to consider the increasingly important ways in which not simply numerical but also social 'group unities' or collectives are attained in the use of Big Data, adding a new dimension to the emerging ethical and legal debates. Such an approach would reinforce the development of a shared literacy by showing its value in terms of the distinctively social implications of the emergence of Big Data.

In (temporary) conclusion, we believe that a social framework for Big Data that draws on both these senses of 'socializing', identified through the collaborative process we have described here, will have the benefit of being able to direct and inform the capacity of Big Data to socialize for the public good.

---

[8] An example of the new kinds of understanding to be derived from the approach is McNally, R. and Mackenzie. A. (2012) 'Understanding the 'intensive' in data intensive research: data flows in next generation sequencing and environmental networked sensors', *The International Journal of Data Curation*, 7(1).

[9] Whitehead, Alfred North (1968). *Modes of Thought.* New York, Free Press.

## References


Beer, David, and Roger Burrows. 2013. 'Popular Culture, Digital Archives and the New Social Life of Data,' Theory, Culture & Society 30, 4: 47-71.

Bowker, G. C. and S. L. Star. 1999. *Sorting Things Out: Classification and its Consequences*. Cambridge, Massachusetts, The MIT Press.

Kopytoff, Igor. 1986. 'The Cultural Biography of Things: Commoditization as Process,' in *The Social Life of Things*, ed. Arjun Appadurai. Cambridge University Press, 64-91.

Lash, Scott and Celia Lury. 2007. *Global Culture Industry: The Mediation of Things*, Cambridge, Polity.

Law, John, Evelyn Ruppert and Mike Savage. 2011. 'The Double Social Life of Methods,' CRESC Working Paper Series, Paper No. 95.

Marcus, George E. (ed.) 2000. *Para-Sites: A Casebook Against Cynical Reason*, Chicago, University of Chicago Press.

McNally, R. and Mackenzie. A. 2012. 'Understanding the 'intensive' in data intensive research: data flows in next generation sequencing and environmental networked sensors', *The International Journal of Data Curation*, 7(1).

Rabinow, P., G. E. Marcus, J. Faubion and T. Rees. 2008. *Designs for an Anthropology of the Contemporary*. Durham, Duke University Press.

Stapleton, L. K. (2011). 'Taming Big Data'. *IBM Data Magazine*. 16: 1-6

The Center for Ethnography. 2009. 'Center as Para-site in Ethnographic Research Projects', University of California, Irvine. http://www.socsci.uci.edu/~ethnog/theme3.htm.

Whitehead, Alfred North. 1968. *Modes of Thought*. New York: Free Press.

## Appendix: Background Summaries on Key Concepts

Digital Data

Big Data

Digital Data-Object

Boundary Object

Collaboratory

## Summary: Digital Data

The ubiquity of digital devices and the data they generate - from that of social media platforms and browsers to those of online purchasing and sensors – and their implications for empirical methods in the social sciences are a matter of some debate. Within sociology, for example, digital data are having an impact on the so-called key evidentiary bases of sociology and leading to a revitalized concern with what 'the empirical is and how it matters' in the discipline (Adkins and Lury 2009: 4). 'As more and more behaviour is conducted electronically, more and more things can be measured more and more often' and this requires that we 'rethink data analysis from the ground up' (Abbott 2000: 298, 299). Because digital data now 'moves, flows, leaks, overflows and circulates beyond the systems and events in which it originates' it is changing both the measures and values of the contemporary world (Adkins and Lury 2009: 4). On the one hand, digital data are said to be challenging the expertise of sociologists in both the generation and analysis of social life, a point advanced by Savage and Burrows (Savage and Burrows 2007). They argue that social science methods are unable to organise 'lively' sources such as 'social' transactional data, which are now routinely collected, processed and analysed by a wide variety of private and public institutions and represent a coming crisis for empirical sociology's jurisdiction for knowing social relations. But new sources of data are not only understood as a crisis but also a provocation to the discipline to invent methods that can adapt, re-purpose and engage with digital media (Adkins and Lury 2009, Back and Puwar 2012).

For Marres (2012) sociological methods have always involved distributions of roles between the academy and other actors (in industry for e.g.) and which are now being redistributed in ways that are more open-ended and reconfiguring. Similarly Ruppert et al. (2013) argue that digital data and devices call for reassembling social science methods and how they remake 'old' techniques (e.g., surveillance) and assumptions about who are the subjects and objects of knowledge.

Digital data are generated by practices that engage, relate to and involve what could be called participatory arrangements where subjects are more active in how data is generated (Marres 2012). For Adkins and Lury, new sources of data are closing a gap between the practices of sociologists and those of social worlds. On the one hand, social media platforms are mediums of digital sociality and the doing of social relations. The data they generate in the cultural sphere on platforms such as Facebook, Spotify and Flickr are also part of everyday popular cultural forms that are actively both produced and consumed via myriad acts of 'playbour' (Beer and Burrows 2013). Such data is lively as it is recursively taken up and re-appropriated as a part of contemporary popular culture. At the same time social researchers and

others develop methods for analysing and interpreting the data these platforms generate to make sense of, interpret and know those digitally mediated lives (Ruppert, Law et al. 2013). Thus digital mediums both open up the possibilities for creative, interactive, and collaborative research engagements with publics and at the same time can render them unknowing research subjects. Their agential capacities are thus variably configured by the specific method relations of which they become a part.

On the other hand, while the rise of participatory user-led Web resources have been associated with 'empowerment' and 'democratisation' (Beer 2009) (Beer and Burrows 2007), data analysis typically involves the use of powerful algorithms (Lash 2007). While not a new phenomenon, the rise in vast amounts of digital data has increased their ubiquity and influence. Predictive modeling and correlations are often used to make causal inferences to categorise subjects (Mayer-Schonberger & Cukier, 2013). The propensity for data predictions to be used by organisations (government, commercial, research) is turning users into subjects and objects of knowledge, and can lead to penalising certain groups on the basis of algorithmic predictions such as in predictive policing and health care (Mayer-Schönberger and Cukier 2013).

**References**


Abbott, A. (2000). 'Reflections on the future of sociology', *Contemporary Sociology* 29(2): 296–300.

Adkins, Lisa, and Celia Lury (2009). 'Introduction to special issue, "What is the empirical?".' *European Journal of Social Theory* no. 12:5-20.

Back, Les, and Nirmal Puwar (2012). A manifesto for live methods: Provocations and capacities.' *The Sociological Review* no. 60 (S1):6-17.

Beer, David (2009). 'Power through the algorithm? Participatory web cultures and the technological unconscious.' *New Media & Society* no. 11 (6):985-1002.

Beer, David, and Roger Burrows (2007). 'Sociology and, of and in Web 2.0: Some Initial Considerations.' *Sociological Research Online* no. 12 (5):1-18.

Beer, David, and Roger Burrows (2013). 'Popular Culture, Digital Archives and the New Social Life of Data.' *Theory, Culture & Society* no. 30 (4):47-71.

Lash, Scott (2007). 'Power after Hegemony: Cultural Studies in Mutation?' *Theory, Culture & Society* no. 24 (3):55-78.

Marres, Noortje (2012). 'The redistribution of methods: On intervention in digital social research, broadly conceived.' *The Sociological Review* no. 60 (S1):139-165.

Mayer-Schönberger, Viktor, and Kenneth Cukier (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. London: John Murray.

Ruppert, Evelyn, John Law, and Mike Savage (2013). 'Reassembling Social Science Methods: The Challenge of Digital Devices.' *Theory, Culture & Society* no. 30 (4):22-46.

Savage, Mike, and Roger Burrows (2007). 'The Coming Crisis of Empirical Sociology.' *Sociology* no. 41 (5):885-899.

## Summary: Big Data

In a very short time what was initially referred to as the 'data deluge' (Hey and Trefethen 2003), information overload or tsunami of data has come to known as 'big data.' While variously defined, Big Data refers to digital content generated either online or offline in social, commercial, scientific, and governmental databases. Though sometimes referred to as simply the latest buzzword or bandwagon and criticized for being substantively vague, it has gained popular salience[10] and this is one among many reasons for adopting and engaging with it (Manovich 2011, boyd and Crawford 2012). Another reason is its increasing use in industry, government (Letouze 2012) and by numerous social science scholars in sociology (Venturini, Jensen et al. forthcoming), anthropology[11], geography (Kitchin 2014; Crampton et al. 2012), journalism, cultural studies and humanities (Manovich 2009, Berry 2011), population studies (Sobek, Cleveland et al. 2011) and in the sciences of biology (Leonelli 2012, Strasser 2012), information (Shiri 2012) and computer science (Lazer et al. 2009 ).

This diverse and far-reaching take up of the term across disciplines is also indicative of the fundamental impact that Big Data is having from reinventing society, transforming notions of identity, influencing government policy-making, mobilising a radical change in information production, changing practices of international development, making governments transparent and more accountable, creating and formatting new economies, changing the very material of scientific inquiry and knowledge and leading to alternative social theories of individuals and societies.

The meaning and relevance of the term is a matter of some debate (Floridi 2012). Some trace its etymology back to the 1990s and to Silicon Graphics, a giant of computer graphics that dealt with new kinds of data such as Hollywood special-effects to video surveillance by spy agencies (Lohr 2013). But as many analysts have noted the existence and processing of large volumes of data is not new. Jacobs

[10] E.g., the Quantified Self and The Human Face of Big Data project. Quantified Self is an initiative for people to share tools and ideas for analysing large quantities of data compiled through self-tracking devices (http://quantifiedself.com/about/). The Human Face of Big Data is a project initiated by Rick Smolan, a former Time, Life, and National Geographic photographer, and creator of the Day in the Life book series. It is a 'globally crowdsourced media project focusing on humanity's new ability to collect, analyze, triangulate and visualize vast amounts of data in real time' (http://humanfaceofbigdata.com).

[11] See for example, see Jenna Burrell 'The Ethnographer's Complete Guide to Big Data: Small Data People in a Big Data World', Available at: http://ethnographymatters.net/2012/05/28/small-data-people-in-a-big-data-world/.

(2009) for example notes that in the 1980s when social scientists gained access to the entire 1980 U.S. Census database—some 100GB of data drawn from datasets of varying sizes—this certainly constituted big data. And Strasser (2012) has noted that life sciences have dealt with the challenges of massive amounts of data since the Renaissance. On the grounds of volume alone, definitions of what constitutes Big Data certainly vary by subject matter and discipline. Industry and natural science definitions may well be considerably different from those of the social sciences. But for most commentators Big Data does not simply refer to *volume* (which can be multi-gigabyte to multi-petabyte and beyond), but also the *velocity* of data generation (the speed of collecting data in 'real time') and the *variety* of data sources and formats (increasing array of data types from audio, video, and image data, and the mixing and linking of information collected from diverse sources) (Stapleton 2011). While much attention is paid to data that is generated on the Internet, there is also much that is generated in closed networks and then sometimes distributed on the Internet such as literary texts and open government data (e.g., over 9000 for data.gov.uk). Much data is also generated via crowdsourced and distributed data collection and then shared (e.g., the Galaxy Zoo online astronomy project). Furthermore, some data remains in myriad corporate and government databases with controlled access (such as transactional and administrative data).[12] These data are collected with varying degrees of conscious participation by contributors and exist under a wide array of ownership and control systems.

But it is these very qualities of digital data—the volume, velocity and variety—to varying degrees that make some of it difficult to process and analyse using traditional data management and processing applications. These qualities are thus driving innovations in data structures, computational capacities, and processing tools and analytics beyond those provided by packages such as qualitative data analysis software like NVIVO or quantitative software such as Statistical Analysis Software (SAS). While SAS made possible complex analytics such as correlation and working with various large data sets, new generations of analytics such as the open source platform Hadoop MapReduce enable distributed processing across clusters of computers that significantly extends these computational capacities beyond a desktop computer. Analytic techniques such as network analysis, machine learning, clustering, topic modelling, latent semantic analysis are rapidly transforming many disciplines, including the social sciences. Moreover the ethos surrounding open

[12] Data on over a billion transactions every year is handled by central government in the UK: http://bit.ly/V6IUvJ.

source software development ensures that new techniques are more widely and freely available. Myriad web and mobile applications also extend analytics to models that can 'learn' by continuously discovering patterns (e.g., Facebook, Google) to those that can mine structured and unstructured data to detect correlations to those that can make connections between varieties of ubiquitous data compiled 'on-the-go' via mobile phones and environmental sensors. Finally, all of these analytics also advance the use of visualisation as an interface for interpreting and presenting findings.

Such computational innovations are not only happening in the social sciences but also in the humanities and biological and physical sciences, as well as in industry and business. Big data constitutes a quantum change in scale, breath and complexity such that some approaches in biology can be understood as a science of information management (Callebaut 2012), computer sciences as social computing,[13] humanities as a form of cultural analytics (Manovich 2007), geography as urban informatics and sociology as computational social science (Lazer, Pentland et al. 2009 ).

Computational analytics, which favour positivist methods and analyses using computer generated algorithms, has led to suggestions that 'raw data' (unmediated) can be 'mined' and aggregated independent of human inquiry to predict and make sense of behaviour (Anderson 2008); a view premised on the realist assumption that objects reflect and discover reality. Despite suggestions that 'raw data' has led to an end of theory, this claim is highly contested by social scientists (Davies 2012; Kitchin 2014; Ruppert 2013; Uprichard 2014). Some suggest that 'raw data is an oxymoron' - always constructed in relation to theoretical assumptions and methods (Bowker 2013, Gitelman 2013).[14] Rather than a call for turning social scientists into computer scientists, their interventions call for 'socialising' what could easily become a positivist science of individuals and societies.

### References

Anderson, C. (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*, 23.06.08, 1-2.

Berry, D. M. (2011) 'The Computational Turn: Thinking About the Digital Humanities', *Culture Machine,* vol. 12, no., pp. 1-22.

---

[13] See for example, Intel Labs ISTC for Social Computing, https://www.intel-university-collaboration.net/exploratory-research/intel-science-and-technology-center-for-social-computing.

[14] Standard modes of scalability and new management technologies are required to turn unstructured data into meaningful information.

Bowker, G. C. (2013) 'Data Flakes: An Afterward to “Raw Data” Is an Oxymoron ', in L. Gitelman (eds) *“Raw Data” Is an Oxymoron,* MIT Press, Cambridge, MA, pp. 167-171.

Boyd, D., and K. Crawford (2012) 'Critical Questions for Big Data', *Information, Communication & Society,* vol. 15, no. 5, pp. 662-679.

Callebaut, W. (2012) 'Scientific Perspectivism: A Philosopher of Science's Response to the Challenge of Big Data Biology.', *Studies in history and philosophy of biological and biomedical sciences,* vol. 43, no., pp. 69-80.

Crampton, Jeremy W, Mark Graham, Ate Poorthuis, Taylor Shelton, Monica Stephens, Matthew W. Wilson, and Matthew Zook (2012) 'Beyond the Geotag? Deconstructing “Big Data” and Leveraging the Potential of the Geoweb,' *Cartography and Geographic Information Science (CaGIS)* 40, 2: 130-139.

Davies, W. (2013) 'Empirical limits.' *RSA.* Issue 4.

Floridi, L. (2012) 'Big Data and Their Epistemological Challenge', *Philosophy and Technology,* vol. 25, no. 4, pp. 435-437.

Gitelman, L., ed. (2013) *"Raw Data" Is an Oxymoron*, MIT Press, Cambridge, MA.

Hey, T., and A. Trefethen (2003) 'The Data Deluge: An E-Science Perspective', in F. Berman, G. Fox and T. Hey (eds) *Grid Computing: Making the Global Infrastructure a Reality,* John Wiley & Sons Ltd., Chichester, pp. 1-17.

Jacobs, A. (2009) 'The Pathologies of Big Data', *Communications of the ACM,* vol. 52, no. 8, pp. 36-44.

Kitchin R. (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, London: Sage.

Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. V. Alstyne. (2009 ) Computational Social Science. *Science*, 6 February, 721-723.

Leonelli, S. (2012) 'Introduction: Making Sense of Data-Driven Research in the Biological and Biomedical Sciences', *Studies in History and Philosophy of Biological and Biomedical Sciences,* vol. 43, no., pp. 1-3.

Letouze, E. (2012) *Big Data for Development : Challenges & Opportunities*. New York: United Nations Global Pulse.

Lohr, S. (2013) 'The Origins of “Big Data”: An Etymological Detective Story'. *The New York Times*, 1 February.

Manovich, L. (2007) 'Cultural Analytics: Analysis and Visualization of Large Cultural Data Sets', *The Chronicle of Higher Education,* vol., no., pp. 3-23.

——— (2009) 'Cultural Analytics: Visualising Cultural Patterns in the Era of “More Media”', *Domus,* vol., no., pp. 1-4.

———. (2011) Trending: The Promises and the Challenges of Big Social Data. Avail. at http://lab.softwarestudies.com/2011/04/new-article-by-lev-manovich-trending.html.

Pentland, A. S. (2012) Reinventing Society in the Wake of Big Dat*a. Edge* 20.

Ruppert, E. (2013). 'Rethinking Empirical Social Sciences'. *Dialogues in Human Geography* 3(3).

Shiri, A. (2012) 'Typology and Analysis of Big Data: An Information Science Prospective'. Presentation at *Internet, Politics, Policy 2012: Big Data, Big Challenges?* Oxford Internet Institute, St Anne's College, Oxford, UK.

Sobek, M., L. Cleveland, S. Flood, P. K. Hall, M. L. King, S. Ruggles, and M. Schroeder (2011) 'Big Data: Large-Scale Historical Infrastructure from the Minnesota Population Center', *Historical Methods,* vol. 44, no., pp. 61-68.

Stapleton, L. K. (2011) Taming Big Data. *IBM Data Magazine*, April 15, 1-6.

Strasser, B. J. (2012) 'Data-Driven Sciences: From Wonder Cabinets to Electronic Databases.', *Studies in history and philosophy of biological and biomedical sciences,* vol. 43, no., pp. 85-7.

Uprichard, E. (2013) 'Big Data, Little Questions?' *Discover Society*. Issue 1.

Venturini, T., P. Jensen, M. Jacomy, S. Grauwin, D. Boullier, and B. Latour (forthcoming) 'From Explanation to Description: On Mapping in Computational Social Sciences'. *Nature*.

## Summary: Digital Data Object

The term ‘digital data object’ (DDO) is generally employed in the computing and information sciences to denote digitally stored data: ‘computer-based, machine-readable resources (such as web pages or electronic journals), whose information content can be stored and accessed independently of the form in which it was originally created’ (Chilvers and Feather 1998: 365).

There are a number of challenges involved with maintaining the intellectual content of DDOs, which, at present, are non-uniform and characterised by interoperability between existing metadata standards (Day 1996, Woodley 2000). Whereas the medium and the message of data objects were traditionally considered inseparable (Hildreth 1996), DDOs make such a separation possible. The ‘new autonomy’ of data (Lash 2002), and their evanescent nature, presents novel management challenges to ensure that such data is authentic and preserved in its original form. The fact that management practices are generally informed by commercial interests raises further issues relating to value (selection criteria), copyright, access and trust (Chilvers and Feather 1998). These problems are confounded by the fact that metadata standards are rapidly changing and the policies to address these issues are in a rudimentary stage of development. The challenge for data managers is to find new analytical resources to cope with the volume (Abbott 2006) and ‘malleability’ of DDOs (Neavill 1984).

Collecting and analysing digital data raises issues of data quality, representation, durability, validity (Graham 1997), data storage, ownership and management (Chilvers and Feather 1998, Chilvers 2002). In this regard, the computing and information sciences generally define DDOs in relation to interoperability, metadata, and management rather than the infrastructures and investments that have gone into making them up. Fuller (2004) describes this as the distributed work activity involved in composing digital objects and the specificities of their contexts.

### References

Abbott, A. (2006) 'Reconceptualizing Knowledge Accumulation in Sociology', *American Sociologist*, vol. 37, no. 2.

Chilvers, A. (2002) 'The Super-Metadata Framework for Managing Long-Term Access to Digital Data Objects: A Possible Way Forward with Specific Reference to the Uk', *Journal of Documentation*, vol. 58, no. 2.

Chilvers, A., and J. Feather (1998) 'The Management of Digital Data: A Metadata Approach', Electronic Library, vol. 16, no. 6.

Day, M. (1996) Interoperability between Metadata Formats, Roads Workshop, Welcome Institute. London: Wellcome Institute. <http://www.ukoln.ac.uk/metadata/presentations/roads-august1996/%3E.

Fuller, M. (2004) 'Digital Objects.' Paper presented at the Read_Me Software Art and Cultures Conference. <http://runme.org/project/+digitalobjects/>

Graham, P. S. (1997) 'Building the Digital Research Library: Preservation and Access at the Heart of Scholarship.' In Follett Lecture Series. Leicester University.

Hildreth, C. R. (1996) 'Preserving What We Really Want to Access, the Message Not the Medium: Challenges and Opportunities in the Digital Age', in A. H. Helal and J. W. Weiss (eds) *Electronic Documents and Information: From Preservation to Access*, Publications of Essen University Library, Germany, pp. 78-95.

Lash, S. (2002) *Critique of Information,* Sage, London.

Neavill, G. B. (1984) 'Electronic Publishing, Libraries and the Survival of Information', *Library Resources & Technical Services*, vol. 28, no. 1.

Woodley, M. (2000) 'Crosswalks: The Path to Universal Access', in M. Baca, P. Harpring, J. Ward and A. Beecroft (eds) *Introduction to Metadata: Pathways to Digital Information*, The J. Paul Getty Trust, Los Angeles.

## Summary: Boundary Object

> How can we see and analyse something so ubiquitous and infrastructural—something so 'in between' a thing and an action? (Bowker and Star 1999: 285)

Bowker and Star develop their understanding of boundary object through an analysis of how formal classification systems seek to regularize the movement of information from one context to another and across time and space. Boundary objects are classifications that manage the tension between multiple interpretations across contexts. The concept recognizes that multiplicity is given and not incidental and is what makes classification and the constitution of the boundary object necessary.

If both people and information objects inhabit multiple contexts simultaneously and if the goal of information systems is to transmit information across these contexts then specific means are required to enable this to happen. The multiple contexts can be understood as different communities of practice/social worlds, that is, as sets of relations among people 'doing things together' (Becker) (material and symbolic) where their activities, routines and practices constitute structures. Being a member includes familiarity with specific categories that apply to encounters with objects and people and deep familiarity with these leads to the naturalization of a community's categories. Membership is thus the experience of common encounters that are increasingly naturalized.

Leigh Star initially coined the boundary object as a way to talk about how scientists do this, how they balance different categories and meaning across contexts (Star and Griesemer 1989). They inhabit several communities of practice and need to satisfy the informational requirements of each. Their concepts must thus be plastic enough to adapt to local needs and robust enough to maintain a common identity across sites. Another way of putting this is that they need to have categories that are 'weakly structured in common use and strongly structured in individual-site use' (Bowker and Star 1999: 297). In this way the boundary object is a 'medium of communication' that can maintain coherence across intersecting communities, be recognizable to each and be simultaneously 'concrete and abstract'.

The boundary object arises over time from durable cooperation among communities of practice. They are working arrangements that resolve anomalies of naturalization without imposing a naturalization of categories from one community or from an outside source of standardization – they are therefore most useful in analyzing cooperative and relatively equal situations rather than impositions. How are boundary objects established and maintained? When a category becomes an object

existing in more than one community then it is a medium of communication. The relationship of a newcomer to a particular context largely revolves around the nature of relations with objects and not, counter-intuitively, directly with the people, that is, the objects mediate relations. The object is naturalized when we strip away its creation and situated nature; members forget its local nature or the actions that maintain and recreate its meaning.

## References


Bowker, G. C., and S. L. Star (1999) *Sorting Things Out: Classification and Its Consequences*, The MIT Press, Cambridge, Massachusetts.

Star, S. L., and J. R. Griesemer (1989) 'Institutional Ecology, "Translations" and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39:' *Social Studies of Science,* vol. 19, no., pp. 387-420.

## Summary: Collaboratory

A collaboratory is a collective mode of inquiry which involves inventing new forms of work that seek to redistribute individual and collective contributions (Rabinow 2006: 1-2). Whereas the term is typically reserved in the natural sciences and computing to denote a distributed research network (Collier in (Rees Instigator (2007): 54)), for those working within the Anthropology of the Contemporary (ARC) model, the term has a distinct meaning:

> A collaboratory is more than an elaborate collection of information and communications technologies. [It is] a new networked organisational form that also includes social processes; collaboration techniques; formal and informal communication; and agreement on norms, principles, values, and rules (Cogburn 2003: 86).

The ARC's model of a collaboratory emerged in response to the so-called 'crisis of method' in American anthropology (Rees & Collier in (Rees Instigator (2007): 2)); namely, dissatisfaction with the individual project model, which emphasises individual achievement, innovation and technique (e.g. ethnography), rather than method (Collier et al. in (Rees Instigator (2007): 10-13)). The collaborative process, conversely, commences from the problem of method – that is, how techniques of data-gathering interact with concept formation and the establishment of collective norms and conventions to produce truth claims and knowledge (Marcus et al. in (Rees Instigator (2007)). Collaboratories aim to create practices of knowledge production, dissemination and critique (Rabinow 2006), and to invent new forms of ethics and writing by reflecting both critically and collectively on the practices and norms of inquiry that orient prevailing discussions of method. From this perspective, method is necessarily collective (Collier in (Rees Instigator (2007)). What constitutes a serious problem and a significant 'finding' can only be defined in a collective context in which topics and objects of study remain open to debate from a variety of stakeholders. In this way consensus emerges through shared standards and critical rectification rather than preconceived truths or established hierarchies.

This collaborative endeavour is referred to as a 'laboratory' and is a critical component of successful experimentation. In contrast to the natural sciences, in the human sciences a laboratory seeks neither to 'discover' positivist truths, nor to generate universal claims about the human condition (Collier et al. in (Rees Instigator (2007): 8)). Instead, it aims to move methodological conversation beyond ethnography by developing collective work on shared problems and concepts. The practical organisation of a collaboratory also differs from a laboratory in the natural sciences in that it is characterised by multi-sited, cross-disciplinary, collective

knowledge making (e.g. regular meetings, co-authored publications), rather than conventional hierarchies or divisions of labour. A collaboratory, then, is distinct from collaboration in the traditional sense of the word in that it produces collective rather than collected work.

On a practical level, a collaboratory involves a rigorous process of concept-formation and experimentation. Initially, the collaborative process requires problematisation - thinking about research questions as problems, and exploring different configurations of inquiry and critique – remaining subject to revision, thereby, favouring experimentation over precision. Concept work plays a central role in this process. It consists in formulating and specifying the meanings of concepts, as well as their capacity to describe research objects. As a practice, collaboratories function as incubators of shared concepts and ideas (Marcus in (Rees Instigator (2007): 35-6)), the aim of which is to invent tools for thought in a mode of collaboration rather than theory. But collaborative work is not just analytic, it is synthetic and recursive, involving a process of reconfiguration and reformulation so as to respond to *emergent* futures with 'preparedness' (Fearnley 2007) and possible solutions.

A collaboratory aims to enhance the social world ethically, politically and ontologically (Rabinow 2006). Politically, the collaborative process interrogates how human life becomes a political problem by examining the practices of experts – the 'styles of reasoning' that experts employ (Hacking 2012). It is premised on the Foucauldian view that investigation should be preceded by examining how objects of knowledge are problematised and produced (Marcus et al. in (Rees Instigator (2007): 22-24)). Analytics and ethics thsu emerge from a problem-space as it unfolds through collaborative engagement (Rabinow and Bennett 2012b). A collaboratory, then, results in both epistemological and ontological 'ramifications'. By disrupting existing hierarchies, and interrogating the sites of power/ knowledge, it consists in re-formulating practices of knowledge production, dissemination and critique, examining how things in the world are constituted as objects (Rabinow and Bennett 2012: 11). It is a pragmatist epistemology (Dewey 2004) that adheres to a social constructionist position, acknowledging that meaning is dynamic and constructed rather than reflecting reality (Rabinow 2007). This emphasis on knowledge production, and the historical contingency of truth claims and practices, is an essential component of ontology because it highlights that alternative modes of being are possible. In addition to contingency, the collaborative mode of inquiry emphasises emergence: developing methods appropriate to the dynamic conditions of contemporary social life. The collaborative process also results in pedagogical outcomes. By rethinking and altering the norms and forms of dissertation training and production (Marcus in (Rees Instigator (2007): 38)), collaborative practices

inform the training process through which students are transformed into scholars (Marcus 2008).

**References**


Cogburn, Derrick L. (2003). 'HCI in the So-Called Developing World: What's in it for Everyone.' *Interactions* no. 10 (2):80-87.

Dewey, John (2004). *Essays in Experimental Logic*. New York: Dover Publications.

Fearnley, Lyle (2007). Pathogens and the Strategy of Preparedness. Anthropological Research on the Contemporary.

Hacking, Ian (2012). ' "Language, Truth and Reason" 30years later ' *Studies in History and Philosophy of Science* no. 43 (4):599-609.

Marcus, George E. (2008). 'Collaborative Options and Pedagogical Experiment in Anthropological Research on Experts and Policy Processes.' *Anthropology in Action* no. 15 (2):47-57.

Pottage, Alain (2014). 'From theory to inquiry?' Review dialogue with Paul Rabinow and Gaymon Bennett. Designing human practices: an experiment with synthetic biology. 2012. Chicago: University Press).

Rabinow, Paul (2006). 'Steps toward an anthropological laboratory.' In *ARC Concept Note*: Anthropological Research on the Contemporary (accessed 02.02.06).

Rabinow, Paul (2007). *Marking Time: On the Anthropology of the Contemporary*. Princeton: Princeton University Press.

Rabinow, Paul, and Gaymon Bennett (2012a). *Contemporary Equipment: A Diagnostic*. In: Anthropological Research on the Contemporary.

Rabinow, Paul, and Gaymon Bennett (2012b). *Designing human practices: an experiment with synthetic biology*. Chicago: University Press.

Rees, Tobias. Instigator (2007). 'Concept Work and Collaboration in the Anthropology of the Contemporary.' In: Anthropological Research on the Contemporary.